

# Digitale Zwillinge als Insight-Quelle

## Potenzial, Fortschritt und Grenzen synthetischer Befragter

Midjourney AI prompt by GCO

### AUTOR:INNEN

#### Carolin Kaiser

Head of Artificial Intelligence  
Nürnberg Institut für Marktentscheidungen

#### Jakob Kaiser

Senior Researcher  
Nürnberg Institut für Marktentscheidungen

#### Rene Schallner

Gründer  
AI Research & Technology Lab

#### Vladimir Manewitsch

Senior Researcher  
Nürnberg Institut für Marktentscheidungen

#### Lea Rau

Doktorandin und  
wissenschaftliche Mitarbeiterin  
Institut für Marketing  
LMU Munich School of Management  
Ludwig-Maximilians-Universität München

### KEYWORDS

Synthetische Umfragedaten  
Large Language Models  
Digitale Zwillinge  
Brand Funnel

Marketingteams brauchen schnelle und verlässliche Einblicke in Einstellungen und Vorlieben von Konsumenten. Klassische Umfragen liefern diese Einblicke jedoch immer schlechter: Sie sind teuer, dauern lange und immer weniger Menschen nehmen daran teil. Deshalb wächst das Interesse an Alternativen, die günstiger sind, sich leichter skalieren lassen und auch unter Zeit- und Budgetdruck Erkenntnisse liefern können. Neue Fortschritte bei großen Sprachmodellen (LLMs) haben einen solchen Ansatz möglich gemacht: Ein LLM beantwortet Fragebögen so, als wäre es ein Konsument mit einem bestimmten demografischen Profil. Auf diese Weise entstehen synthetische Datensätze aus virtuellen Befragten – sogenannte Silicon Samples. Dieser Ansatz findet zunehmend Aufmerksamkeit in der professionellen Marktforschung. Wenn er funktioniert, könnten synthetische Befragte klassische Umfragen ergänzen oder für schnelle Simulationen genutzt werden, etwa um frühe Konsumentenreaktionen oder Konzepte zu testen. Die bisherigen Ergebnisse sind jedoch uneinheitlich, und die meisten Validierungen stammen nicht aus dem Marketing. Es ist noch unklar, inwieweit LLMs tatsächliche Konsumentenentscheidungen, Stimmungen und realistische Unterschiede zwischen Personen zuverlässig abbilden können. Wir stellen eine Studie mit synthetischen Befragten vor, in der zwei ►

## Vergleich der Antworten von echten und synthetischen Befragten unter Verwendung zweier unterschiedlicher Ansätze

BOX 1

### Fokus der Studie: Wahrnehmung von Softdrink- Marken

Im Mittelpunkt der Studie steht die Wahrnehmung von Softdrink-Marken durch Konsumenten – eine Produktkategorie, mit der US-Konsumenten sehr vertraut sind. Wir entwickelten einen Fragebogen zu acht Marken: vier sehr bekannten (Coca-Cola, Pepsi, Sprite, 7UP) und vier deutlich weniger bekannten (Dry, Moxie, Blue Sky, Orangina). So konnten wir prüfen, ob sich synthetische Daten für Massenmarken und Nischenmarken unterscheiden.

### Reale Daten als Benchmark

Als Referenz dienten reale Umfragedaten aus einer repräsentativen US-Stichprobe. Befragt wurden 461 Konsumenten, die regelmäßig Softdrinks kaufen. Die Online-Stichprobe wurde nach Alter, Geschlecht und ethnischer Zugehörigkeit an die US-Bevölkerung angepasst. Diese realen Antworten bildeten die Benchmark für den Vergleich mit den Antworten der digitalen Zwillinge.

### Synthetische Stichprobe: Digitale Zwillinge realer Personen

Für jede reale Person erstellten wir einen digitalen Zwilling, der ihr demografisches Profil abbildete – darunter Alter, Geschlecht, Bildung, Beruf, finanzielle Situation, Wohnort, ethnische Zugehörigkeit und Sprache. Das LLM beantwortete anschließend exakt dieselben Fragen wie die jeweilige reale Person. So konnten wir synthetische und reale Antworten direkt vergleichen – sowohl entlang des Marketing-Funnels als auch bei der Bewertung der Marken. Zum Einsatz kam GPT-4o, das zum Zeitpunkt der Datenerhebung leistungsfähigste verfügbare Modell, mit Standardparametern über die API.

### Messung: Markenauswahl und Markenbewertung

Reale Teilnehmer und digitale Zwillinge beantworteten Fragen zu Markenbekanntheit, Markenberücksichtigung und Kauf. In jeder Phase wählten sie relevante Marken aus. Für jede ausgewählte Marke folgten vier Bewertungsfragen zur Markeneinstellung, etwa zur wahrgenommenen Qualität oder zur Weiterempfehlungswahrscheinlichkeit, gemessen auf einer 7-stufigen Likert-Skala.

### Erweiterte Analyse: Textbasierte statt numerischer Antworten

Da Sprachmodelle besser mit natürlicher Sprache als mit numerischen Werten umgehen können, testeten wir einen alternativen Ansatz. Statt direkt über die Likert-Skala zu bewerten, formulierte das Modell zunächst eine natürliche Textantwort, die einer echten Konsumentenaussage ähnelt. Beispiel: Auf die Frage nach der Weiterempfehlungswahrscheinlichkeit:

unterschiedliche Befragungsformate in einem realistischen Marketing-Funnel-Szenario getestet wurden (Box 1).

### Wie gut sind synthetische Befragte im Vergleich zu echten Personen?

Insgesamt sagten die synthetischen Antworten individuelle Entscheidungen deutlich besser voraus als es bei reinem Raten zu erwarten wäre. Allgemeine Trends – etwa die stärkere Bevorzugung bekannter Marken – wurden korrekt erkannt. Gleichzeitig zeigten sich klare systematische Abweichungen: Die synthetischen Befragten bewerteten Marken insgesamt zu positiv, besonders bekannte Marken, und unterschieden sich untereinander deutlich weniger als echte Menschen.

### Richtige Trends, überhöhte Wahrscheinlichkeiten und Bewertungen

Wie erwartet wählten echte Teilnehmer im Marketing-Funnel deutlich häufiger bekannte als unbekanntere Marken. Dieses Grundmuster wurde von den synthetischen Daten korrekt nachgebildet. Allerdings überschätzten die synthetischen Befragten die Auswahlwahrscheinlichkeit bekannter Marken deutlich (siehe Abbildung 1). Im Durch-

schnitt stimmten synthetische und reale Entscheidungen in 79 % der Fälle überein. Das zeigt: LLM-basierte Befragte erfassen die Richtung der Effekte, überschätzen aber systematisch, wie stark Konsumenten sich tatsächlich für eine Marke entscheiden. Ein ähnliches Bild ergab sich bei den Markenbewertungen. Echte Teilnehmer bewerteten bekannte Marken besser als weniger bekannte, und auch hier trafen die synthetischen Daten den Trend. Die absoluten Bewertungen fielen jedoch durchweg zu positiv aus – mit der stärksten Verzerrung zugunsten bekannter Marken. Über alle Marken hinweg wichen die synthetischen Bewertungen auf der 7-stufigen Skala im Schnitt um 1,2 Punkte von den realen Bewertungen ab.

### Geringere Vielfalt als bei echten Antworten

Durchschnittswerte sind wichtig, aber für Marktforschung ist auch entscheidend, wie stark sich einzelne Konsumenten in ihren Meinungen unterscheiden. Genau hier zeigen sich klare Grenzen synthetischer Daten.

Beim Vergleich der Antwortverteilungen – sowohl bei der Markenauswahl als auch bei den Markenbewertungen auf der Likert-Skala – waren die synthetischen Antworten deutlich weniger streuend als die realen Antworten. Dieser Effekt war besonders ausgeprägt bei bekannten Marken.

wahrscheinlichkeit könnte das Modell antworten: „Die Marke X ist eine solide Wahl, deshalb würde ich sie wahrscheinlich weiterempfehlen.“ Für jede Stufe der Likert-Skala erstellten wir Referenzaussagen, die typische sprachliche Formulierungen für diese Bewertung widerspiegeln. Anschließend wurden semantische Ähnlichkeitsbewertungen durchgeführt und berechnet, welcher Referenzaussage die generierte Antwort sprachlich am meisten entsprach. Die entsprechende Skalenstufe wurde dann als finale Bewertung übernommen. Eine Bewertung von 5 auf einer 7-Punkte-Likert-Skala könnte beispielsweise Aussagen wie „Ich würde dieses Unternehmen wahrscheinlich weiterempfehlen, bin aber nicht ganz überzeugt“ umfassen.



Midjourney AI prompt by GCO

“

Die synthetischen Befragten bewerteten Marken insgesamt zu positiv, besonders bekannte Marken, und unterschieden sich untereinander deutlich weniger als echte Menschen.

”

Insgesamt zeigt sich: Von LLMs erzeugte Daten bilden zwar typische Antworten ab, unterschätzen aber die tatsächliche Meinungsvielfalt realer Konsumenten. Die Ergebnisse fallen homogener aus, als sie es in der Realität sind.

#### **Semantische Ähnlichkeitsbewertungen verbessern den Realismus**

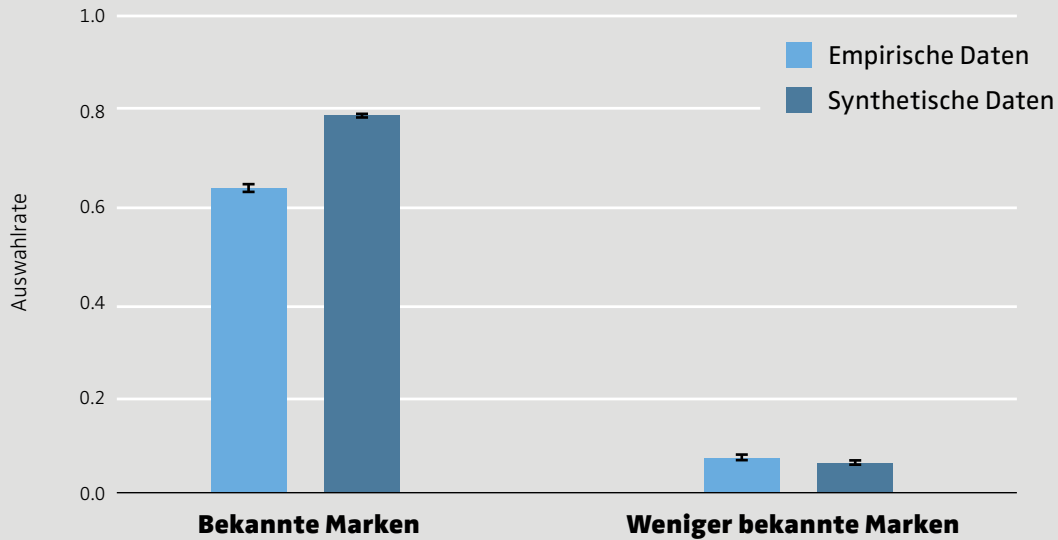
Der Vergleich der Verteilungen der Markenbewertungen von echten Befragten mit synthetischen Befragten auf Basis der direkten Erhebung auf der Likert-Skala und der semantischen Ähnlichkeitsbewertung zeigt für unsere Umfrage ein klares Muster: Die semantische Ähnlichkeitsbewertung erzeugt eine größere Variabilität als die direkte Erhebung auf der Likert-Skala, was auf einen verbesserten Realismus hindeutet (siehe Abbildung 3). Allerdings werden die Markeneinstellungen immer noch überschätzt und es zeigen sich nach wie vor weniger Schwankungen als bei echten Befragten. Beide synthetischen Methoden unterschätzen die in menschlichen Antworten beobachtete Variabilität, obwohl die semantische Ähnlichkeitsbewertung im Vergleich zur direkten Erhebung auf der Likert-Skala eine teilweise Verbesserung zeigt. Kurz gesagt: Die semantische Ähnlichkeitsbewertung ist eine Verbesserung, aber auch noch kein Ersatz für echte Daten.

#### **Implikationen für die Marktforschung: Sinnvolle Einsatzbereiche synthetischer Daten**

Unsere Studie zeigt, dass LLMs mithilfe personalisierter „digitaler Zwillinge“ menschliche Umfrageantworten simulieren können. Der Ansatz hat Potenzial, aber auch klare Grenzen. Synthetische Daten neigen dazu, Antworten zu verallgemeinern oder sozial erwünschte Muster zu zeigen. Dadurch spiegeln sie die feinen Nuancen und die Vielfalt echter Konsumentenmeinungen nur eingeschränkt wider. Der praktische Nutzen liegt der- ▶

## Vergleich der Markenauswahl für bekannte und weniger bekannte Marken – reale vs. synthetische Befragte

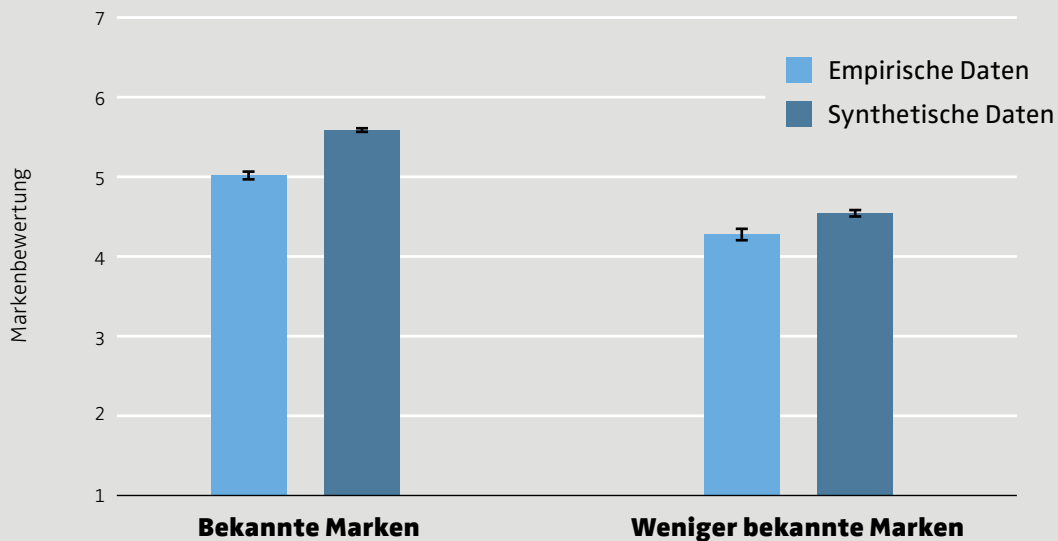
ABBILDUNG 1



Überschätzung bekannter Marken durch synthetische Befragte im Vergleich zu realen Konsumenten.

## Vergleich der Markenbewertungen für bekannte und weniger bekannte Marken – reale vs. synthetische Befragte

ABBILDUNG 2



Synthetische Befragte bewerten Marken systematisch positiver als reale Konsumenten.

zeit vor allem in Frühphasen-Tests, etwa für Konzept- oder Ideenbewertungen, oder in Anwendungen mit geringerem Risiko. Für Entscheidungen, die präzise Einblicke erfordern, sind echte Umfragedaten weiterhin unverzichtbar.

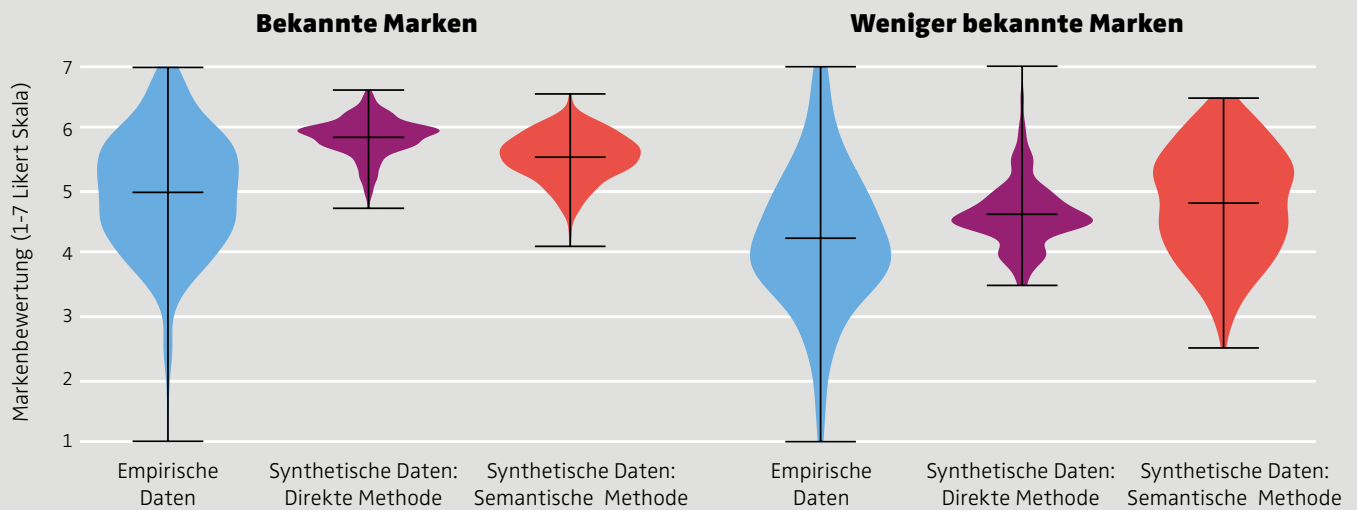
### Ausblick: Wege zur Verbesserung synthetischer Umfragen

LLMs entwickeln sich schnell, und neue Methoden könnten die derzeitigen Einschränkungen synthetischer Umfragedaten verringern. Ein vielverspre-

chender Ansatz ist die hier präsentierte semantische Ähnlichkeitsbewertung (siehe Box 1), bei der das Modell zunächst eine natürliche Textantwort erzeugt, die anschließend auf eine Likert-Skala übertragen wird. So wird die Stärke des Modells in der Textgenerierung genutzt und die Beurteilung ähnelt mehr der durch Menschen. Weitere Verbesserungen könnten erreicht werden, indem mehr persönliche Informationen einbezogen werden – zum Beispiel psychologische Merkmale oder Verhaltensdaten. Externe Kontexte wie Produktbewertungen oder Wirtschaftsindikatoren können

## Verteilungen der Markeneinstellungen bei realen und synthetischen Stichproben – direkte Likert-Antworten und semantische Ähnlichkeitsbewertung

ABBILDUNG 3



**Synthetische Bewertungen sind variabler mit semantischer Ähnlichkeitsbewertung, bleiben aber homogener als echte Antworten.**

“  
Der praktische Nutzen liegt derzeit vor allem in Frühphasen-Tests, etwa für Konzept- oder Ideenbewertungen, oder in Anwendungen mit geringerem Risiko.  
”

auch über Retrieval-Augmented Generation (RAG) inkludiert werden. Insgesamt könnten diese Ansätze dazu beitragen, dass synthetische Daten realistischer werden und die tatsächlichen Einstellungen von Konsumenten besser widerspiegeln.

### Effizienz und Ethik in Einklang bringen

KI bietet Marktforschern viele Chancen: Sie kann Forschungsprozesse beschleunigen, Kosten senken und große Datenmengen schnell generieren. Gleichzeitig besteht das Risiko, dass Verzerrungen

und Übergeneralisierungen die Ergebnisse verfälschen und dann falsche strategische Entscheidungen getroffen werden. Für Konsumenten kann KI ein schnelleres, reaktionsfähigeres Marketing ermöglichen, doch zu starke Vereinfachungen könnten dazu führen, dass Angebote weniger personalisiert oder sogar irrelevant werden. Auf gesellschaftlicher Ebene wirft der zunehmende Einsatz von KI in der Marktforschung auch ethische Fragen auf: Sie kann Stereotype verstärken, Minderheitenmeinungen marginalisieren und traditionelle Aufgaben in der Forschung massiv verändern. Um verantwortungsbewusst Nutzen aus KI zu ziehen, sind daher sorgfältige Kontrolle und neue Kompetenzen erforderlich – damit Erkenntnisse zuverlässig sind, ohne Vorurteile oder Ungleichheiten zu verstärken. ◀

### LITERATURHINWEISE

Kaiser, C., Kaiser, J., Manewitsch, V., Rau, L., & Schallner, R. (2025).

Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (pp. 82–86). Association for Computing Machinery. <https://doi.org/10.1145/3708319.3733685>

Maier, B. F., Aslak, U., Fiaschi, L., Rismal, N., Fletcher, K., Luhmann, C. C., Dow, R., Pappas, K., & Wiecki, T. V. (2025).

LLMs reproduce human purchase intent via semantic similarity elicitation of Likert ratings. arXiv:2510.08338. <https://arxiv.org/abs/2510.08338>